DOCUMENT RESUME

ED 389 741 TM 024 370

AUTHOR	Mills, Craig N.; Stocking, Martha L.
TITLE	Practical Issues in Large-Scale High-Stakes Computerized Adaptive Testing.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-95-23
PUB DATE	Aug 95
NOTE	31p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; Item Banks; Psychometrics; *Scoring; Selection; Simulation; Test Construction; *Testing Problems; *Test Items; Theory Practice Relationship
IDENTIFIERS	*High Stakes Tests; Item Exposure (Tests); *Large Scale Assessment; Test Security

ABSTRACT

Computerized adaptive testing (CAT), while well-grounded in psychometric theory, has had few large-scale applications for high-stakes, secure tests in the past. This is now changing as the cost of computing has declined rapidly. As is always true where theory is translated into practice, many practical issues arise. This paper discusses a number of such issues and outlines potential problems and potential solutions. The purpose is to encourage further development of solutions to these and other practical issues facing measurement professionals involved in CAT implementation. Practical issues explored include: (1) establishing test designs through simulation studies; (2) selecting the first item; (3) choosing the stopping rule; (4) scoring adaptive tests; (5) incorporating the test plan; (6) administering items belonging to sets; (7) facilitating item ordering; (8) controlling item exposure; (9) controlling item overlap; (10) providing item review; (11) dealing with item omissions; (12) allowing for incomplete tests; (13) establishing comparability; (14) developing CAT pools; (15) maintaining CAT pools; (16) developing and prototyping new item types; (17) complying with disclosure requirements; and (18) accommodating low-volume tests. (Contains 19 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-95-23

ED 389 741

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PRACTICAL ISSUES IN LARGE-SCALE HIGH-STAKES COMPUTERIZED ADAPTIVE TESTING

Craig N. Mills
Martha L. Stocking

BEST COPY AVAILABLE

Educational Testing Service
Princeton, New Jersey
August 1995

TM024370

PRACTICAL ISSUES IN LARGE-SCALE HIGH-STAKES COMPUTERIZED ADAPTIVE TESTING

Craig N. Mills
and
Martha L. Stocking

August, 1995
Educational Testing Service
Princeton, NJ 08541

Note: Authors' names are listed alphabetically.

Copyright © 1995. Educational Testing Service. All rights reserved.

Abstract

Computerized adaptive testing (CAT), while well-grounded in psychometric theory, has had few large-scale applications for high-stakes, secure tests in the past. This is now changing as the cost of computing has declined rapidly. As is always true at such junctures where theory is translated into practice, many practical issues arise that must now be addressed. This paper discusses a number of such issues and sketches out potential problems and potential solutions. Our purpose is to encourage further development of solutions to the issues presented as well as other practical issues facing measurement professionals involved with the implementation of adaptive testing.

Introduction

Computerized adaptive testing (CAT; Lord, 1980; Wainer, Dorans, Flaugher, Green, & Mislevy, 1990) is widely recognized as a powerful measurement tool that has the potential to improve testing in a number of ways. It offers the potential to both shorten tests and improve their accuracy. Tailoring tests to test-takers' levels of proficiency may decrease the effect of irrelevant variance (e.g. lower frustration for lower ability students who typically take tests consisting largely of items too difficult for them). Testing may become more convenient since tests may be offered on a much more frequent basis than is typically the case for traditional paper-and-pencil tests. Furthermore, the migration of tests to computer delivery, by itself, offers the potential of supporting a wide variety of new types of assessment which may allow not only more comprehensive assessment of the constructs measured by traditional tests, but also allow additional constructs to be tested efficiently.

Despite the appeal of adaptive testing, practical applications have been limited. Although some low stakes but high volume tests have been made into CATs (for example, Ward, 1988) and there have been limited applications of CAT in some certification settings (see, for example, Lunz & Bergstrom, 1991), until recently, no large-scale, secure, high-stakes tests have been offered predominantly in a computerized adaptive format. In the last year, however, that has changed. Starting with the introduction of a computerized adaptive General Test in 1993 sponsored by the Graduate Record Examination Board and the NCLEX/CAT in 1994 sponsored by the National Council of State Boards of

Nursing, a move to large-scale adaptive testing for secure, high-stakes, tests has clearly begun.

As is true any time theory is translated into practice, the operational implementation of this type of large-scale adaptive testing has raised numerous practical issues that need to be addressed. The purpose of this paper is to discuss some of those issues, and encourage further development of solutions to these and other practical issues facing measurement professionals considering adaptive tests.

Adaptive Testing

At a simple level, adaptive testing is an assessment process whereby the test is constructed as a test-taker is responding to items. Items are selected from a very large, statistically cohesive (unidimensional) item pool based on the test-taker's responses to all previous items. In the simplest form of adaptive testing, if the test-taker misses the first question, an easier question is administered. Conversely, if the first item is answered correctly, a harder question is answered. As the tests progresses, the computer evaluates all previous answers and the statistical characteristics of the items remaining in the pool to identify the 'best' next question. This process continues until some stopping rule is satisfied at which time a score is generated. This 'bare bones" description of a simple adaptive testing item selection rule is summarized below:

1. Select the first item with best measurement at some level of proficiency.
2. If answered correctly, select the next item with best measurement at a high proficiency level. If answered incorrectly, select the next item with best measurement at a low proficiency level.

Continue until there is at least one correct and one incorrect answer.

3. If there is at least one correct and one incorrect answer, estimate proficiency and select the best item for this estimated proficiency.
4. Repeat Step 3 until the stopping rule is satisfied.

This simple model has driven much of the research which provides the foundation for adaptive testing. However, practical applications of adaptive testing are much more complex. For example, even though the items comprising the pool are statistically unidimensional, most test specifications incorporate logical subdivisions of the content, a variety of item formats, and other considerations. For adaptive testing to gain widespread acceptance, it must also incorporate these factors.

Practical Issues

Among the practical issues that guide the development of adaptive tests are:

Establishing test designs through simulation studies,

Selecting the first item,

Choosing the stopping rule,

Scoring adaptive tests,

Incorporating the test plan,

Administering items belonging to sets,

Facilitating item ordering,

Controlling item exposure,

Controlling item overlap,

Providing item review,

Dealing with item omissions,

Allowing for incomplete tests,

Establishing comparability,

Developing CAT pools,

Maintaining CAT pools,

Developing and prototyping new item types,

Complying with disclosure requirements, and

Accommodating low volume tests.

In the remainder of this paper, each of these issues is expanded upon with a discussion of various possible solutions to problems presented.

Establishing test designs through simulation studies

In conventional test construction, tests are carefully constructed and reviewed in advance of being administered. Information from the pretesting of items may be available to aid the test specialist in the test construction process as well as to predict the statistical properties of a test when it is actually administered to test-takers. But the foundation of conventional test assembly is the detailed and meticulous attention by human test specialists.

In contrast, adaptive tests are assembled as test-takers respond to items. Therefore the design of the adaptive test by test specialists must be supported in advance of testing by the creation of tools that are not necessary (although they exist) in the construction of conventional tests.

The modern development of adaptive testing rests on Item Response Theory (IRT; Lord, 1980). By employing the methods available from IRT, it is possible to simulate the responses of typical test-takers to items for which IRT-based item information has been developed from pretesting. Using these

types of tools, it becomes possible to try out many different adaptive test designs and to understand how each design might function in terms of domain coverage and statistical properties down to any desired level of detail, in advance of the administration of any adaptive test to real test-takers. The consequences of decisions made about many (but not all) of the issues listed above are possible to investigate in this manner (see, for example, Eignor, Way, Stocking, and Steffen, 1993).

Of course, establishing adaptive test designs through simulation studies is no substitute for verifying the behavior of the adaptive test with real information from real test-takers. But if the simulation studies are carefully crafted, real information will confirm the results of the simulation studies.

Selecting the First Item

Conventional tests typically begin with the easiest items. This allows most test-takers to begin the test with a "success experience." Although there is no psychometric evidence that an initial success is important to good measurement, test developers often posit that starting the test with more difficult questions may raise anxiety levels and negatively affect performance. However, from a psychometric perspective, adaptive tests will be more efficient if the initial item is of middle difficulty since, in the absence of prior information about a test-taker's proficiency, the best estimate of her proficiency is the average proficiency. The theory does not require an initial item of middle difficulty, so the choice of the first item may be dictated by other considerations.

Many testing programs may want to continue the practice of providing an initial success experience to the majority of test-takers. In that case the

first item is selected to be slightly easier than average. A potential drawback to this approach is that the number of items having an appropriate level of difficulty may be somewhat limited. It may be possible to include a very large number of moderately easy items in the pool to overcome this security concern. While this approach has the advantage of retaining the early likelihood of success for many test-takers, it may result in an item pool that is less than optimal since an overabundance of easier items may be beneficial only in terms of the first item to be administered.

Choosing the Stopping Rule

Conventional tests present all test-takers with the same number of items. Individuals may or may not respond to all questions either because they omit items that they perceive to be too difficult for them or because the test is speeded for them. A disadvantage of fixed length tests is that they provide different levels of score precision at different points on the proficiency scale. Theoretically, adaptive tests can provide equal precision at all points on the scale through the use of variable length tests, although fixed length adaptive tests can also be administered.

Stocking (1987) has shown that variable length testing can result in biased estimates of proficiency if the test is short. This may not be a problem if the number of items to be administered is large (say, more than 50 items) or if the purpose of the test is to make a classification decision (or both). However, if the test score is reported on a continuous scale and the number of questions is small, fixed length adaptive tests are to be preferred.

Scoring Adaptive Tests

Conventional standardized tests are typically scored by transforming the raw score to a scaled score through a process called test equating. This

scaled score is the test score that is comparable to scores earned on different editions of the same test administered at different times. The raw score on a test may be based on the number of items answered correctly (number-right scores) or it may involve a correction to account for the possibility that some correct answers were attained by guessing (formula scores). The same approach may also be applied to adaptive testing (Stocking, 1994; Potenza, 1995). More typically, CATs are scored either with a final proficiency estimate based on the underlying psychometric model, or the proficiency estimate referenced to a number right score on a conventional test or some other collection of items (Ward, 1988; Wainer, et al., 1990).

Although a proficiency estimate based on the underlying psychometric model is typically used in the item selection process in adaptive testing, using such a proficiency estimate as a test score can raise other issues, depending upon the particular model used. It is clear that an individual who answers about half the items on an easy test correctly should receive a lower score than one who answers half the items correctly on a harder test. However, since the adaptive test considers all item responses in generating a score, functionally two test-takers who answer the same item correctly, but whose tests differ in overall difficulty may receive different amounts of credit for that correct answer (depending on the model underlying the test). Since the issue is one that is of concern to the general public and difficult to explain non-technically, programs may wish to consider a scoring model that is different from the item selection model.

Incorporating the test plan

All tests are developed to meet a variety of specifications. Test blueprints incorporate considerations such as variation in content, item

types, representation of important subpopulations, and so forth. Traditional testing programs often have highly detailed specifications that require each item to fulfill multiple components of the specifications. This is important if we are to assert that all test-takers are taking tests of the same construct. It is important that adaptive tests also fulfill the test blueprint. However, in the development of an adaptive test, there may be some trade-offs in terms of determining which specifications are most important and which can be relaxed to take advantage of the power of the measurement models.

Consider, for example, a test of verbal reasoning. Such a test may contain a variety of item formats including reading comprehension items in which a series of questions are based on a stimulus passage. Questions may assess a variety of components of reading comprehension such as identifying the main idea, drawing appropriate inferences, and identifying style and tone. Since an adaptive test will include fewer questions, it may not be possible to include questions on each of these component skills. In determining whether or not the specifications can be relaxed, it is important to relate the specifications to the construct being measured. While the specific skills may be critical to the successful measurement of reading comprehension, it is not clear that they are equally important in the assessment of verbal reasoning. In the latter case, simply assessing a sample of reading comprehension may be sufficient. Other specifications, such as ensuring that the reading passages represent a broad spectrum of content may, however, be critical since failure to include different content may advantage individuals with expertise in the areas that are presented and disadvantage those who do not have such expertise.

It is typically the case that there are many more specifications to be met than there are items to administer. Each item is classified on multiple dimensions of the specifications and, as a result, contributes to the fulfillment of multiple requirements. The adaptive testing algorithm must maintain a record of the extent to which the test in progress meets each specification and select items in a way that best fulfills the specifications. Since all specifications are not equally important, different weights can be applied so that there is greater confidence that the final test meets the most important specifications and deviations, if any, occur for the less important specifications (Stocking & Swanson, 1993; Swanson & Stocking, 1993). This is similar to the trade-offs made by human test developers working with limited item pools.

Administering Items Belonging to Sets

Many tests include items belonging to sets. To continue the previous example, a reading passage may be followed by several items. In conventional tests, all items related to a particular stimulus are presented together. This is a convenience since the stimulus material need only be printed once, but is also important since it would be unfair to require test-takers to access the stimulus material multiple times. Also, given that all test-takers receive the same questions, there is no need to separate them from one another.

From a purely psychometric perspective, it would be best to allow sets to be administered multiple times, if the algorithm selects them. However, the result could be either (1) that an test-taker works through a long stimulus only to answer one question or (2) the same stimulus could be presented many times, but with other items administered in between. Both seem

unfair and, in terms of the time it takes to process the stimulus information, inefficient from the test-taker perspective. Therefore, it is typical for adaptive tests to contain specifications about the number of items to be administered with a stimulus and for all of them to be administered the first time the stimulus is selected. Thus, if there are ten items related to a specific stimulus and the specifications call for three of them to be administered, once the stimulus is selected, the three items (from the ten available) that provide the best measurement would be administered even if there are other items not associated with the stimulus available for administration that would be better at that point in the test.

Facilitating Item Ordering

Conventional tests typically follow a rigid structure for the presentation of items. All items of a certain type may appear together or there may be, for example, three items of Type A followed by four of Type B followed by three more of Type A. There are often both conceptual and practical issues associated with this ordering. Conceptually, it may be important to the measurement of the construct that test-takers be able to develop an orientation to an item type and maintain that orientation throughout a set of questions. Alternately, for some tests, the ability to switch orientation from item to item may be important. Practically, it may be that presenting a block of similar items together provides economies in the printing of the test books.

Adaptive tests typically present only one item on the computer screen at a time. In this case, the directions for responding to the item may become part of the actual item. The consistency of orientation may also be less important since each item is more clearly an independent task. Therefore,

there can usually be more flexibility in the ordering of items than is true for conventional tests. To the extent this is possible, it is psychometrically desirable since it allows measurement considerations to have greater weight than if the selection of the next item is largely dependent on the type of item last administered.

Controlling Item Exposure

Item exposure refers to the rate at which items are presented to test-takers. It includes the percentage of test-takers who see an item, the absolute number who see it, and the period of time during which the exposure occurs. Conventional tests control item exposure by administering small numbers of items to large groups of test-takers. If a testing program administers tests on, for example, four dates per year and a different test is used on each date, twenty-five to thirty percent of the test-takers (several hundred thousand) may see the same questions. However, since they all see the questions on the same day and if the test form is not administered again within the same year, this high rate of exposure may not present a large security risk.

In adaptive testing, the number of test forms becomes very large. However, different forms will contain some of the same questions. Since the tests are typically administered to small groups of test-takers at frequent intervals (as often as every day), all items may be exposed within a very short period of time. Even though a small number of test-takers will have seen each item, there is a risk that the items will become known.

There are several tactics that may be used to control item exposure. The first is to limit exposure through modifications to item selection. The computer might, for example, identify not only the "best" next item, but the

"best N" next items. The item actually selected for administration might be selected from the "best N" based on some mathematical model. A simple random selection may be used (see, for example, McBride & Martin, 1983) or a more complex probabilistic model can be applied (see for example, Sympson & Hetter, 1985; Stocking and Lewis, 1995a; Stocking and Lewis, 1995b). Such attempts to decrease the exposure of items may reduce the psychometric efficiency of the adaptive test, resulting in the requirement of longer test lengths. And the success of attempts to control item exposure will, of course, be effected by the number of items available.

Controlling Item Overlap

Item overlap refers to the extent to which one item may cue the correct response to another item or the extent to which two items depend on the same specific knowledge. For example, the word "panegyric" might be very useful in assessing certain vocabulary skills, but including it twice in a test might be unfair. Conventional tests ensure that overlap is controlled through review of intact tests before they are printed. In the test assembly process, test developers can ensure that overlapping items appear in different tests.

It is likely to be appropriate for adaptive test item pools to contain overlapping items since they may provide measurement at different points in the scale, represent differ types of test questions, etc. However, it is not appropriate for such items to appear together in tests actually administered. Thus, the item selection process must have a mechanism for identifying overlapping items. Once one item from an "overlap group" has been administered, the remaining items in the group should be blocked from administration to the same test-taker.

Providing Item Review

When a test-taker takes a conventional test, she has a limited opportunity to review her answers to previous questions. Review can occur for a variety of reasons. For example, test-takers often, after completing a section of the test, use the time remaining to review their answers to the questions in the section. Review might also occur as a result of a test-taker remembering some fact that she had not previously recalled and recognizing that a previous answer was incorrect. Finally, review can be the result of an item occurring at one point in a test cuing the correct answer to a previous item (obviously, the test development process is designed to ensure that this does not occur). Test-takers do not, however, have unlimited review opportunities. In tests that have separately timed sections, for example, review is limited to the currently administered section.

Most adaptive tests do not permit test-takers to return to previously answered questions. From a theoretical perspective, changing answers to previous questions degrades the psychometric properties of the test and results in a longer test to attain the same measurement accuracy, although these effects may be small for some tests (Lunz, Bergstron, & Wright, 1992). However, more importantly, review presents a fairness issue under some psychometric models. A clever test-taker might use review to "game" the test. In the most extreme case, a bright test-taker might deliberately answer every item presented incorrectly resulting in an adaptive tests consisting of the easiest possible set of questions. Once all items have been presented, the test-taker could review the test, change all answers to correct ones, and, as a result, receive an inappropriately high score.

It is not clear that review is necessary. In fact, research on the issue of comparability of scores between adaptive and linear tests suggest that it is not (Schaeffer, Steffen & Golub-Smith, 1993). However, the lack of review capability is of some concern to test-takers. It remains to be seen whether research will provide a theoretically sound and practically feasible mechanism for incorporating review in adaptive tests.

Dealing With Item Omissions

In conventional tests test-takers are not required to answer every question. Presumably, test-takers omit questions which are too difficult for them. In adaptive tests there should be few such items since most items are closely matched to the test-taker's proficiency. While it is possible to permit omitting in adaptive tests, there are at least three problems associated with doing so.

First, it has long been speculated that different people of similar proficiency omit at different rates. The rate of item omission is likely to be unrelated to the construct being measured, but rather is a reflection of individuals' propensity to take risks. Since this is irrelevant variance for most tests, prohibiting item omissions eliminates this source of error in the scores.

Second, omitting items could also be used as a strategy to manipulate the test unfairly. If an omitted item is not scored, test-takers could simply omit items until they were administered items that were confident they could answer correctly. Thus, they could obtain an artificially high score by answering only those items (see, for example, Lunz & Bergstrom, 1994). Of course, omitted items could be scored as incorrect (which is functionally the case with omitted items on conventional tests where the raw score is the

number of correct answers), but this would not address the issue of differential rates of risk-taking mentioned above.

Finally, allowing omissions in adaptive tests permits unethical test-takers to read through an entire item pool, thus increasing the exposure risk of all items. The compromise of entire item pools, of course, raises the cost of testing for all test-takers.

For these reasons, item omissions are not allowed in most adaptive tests.

Allowing for Incomplete Tests

Despite the fact that most tests are designed to be unspeeded power tests, operational considerations typically require that some time limits be established. Whenever time is limited, some test-takers will not complete the test. A decision must be made whether or not to score an incomplete test and, if so, how much of the test must be completed before the score is generated.

Incomplete tests have not been seen as an issue in conventional tests since, with conventional scoring, if a test-taker answers only a small number of questions, a low score will result. Thus, even though the error associated with that score is very large, it is in the test-taker's interest to answer as many items as possible since correct answers always result in score increases and incorrect answers do not lower scores on conventional (rights scored) tests. Many conventional testing programs will report scores based on as few as a single response.

Reporting scores based on a small number of responses is not feasible with adaptive tests since it is would be in the test-taker's best interest to answer only one question if scores were reported for a one-item adaptive tests AND the test-taker was certain the answer provided was correct. However, it

seems unfair to refuse to report a score if an test-taker has answered all but, for example, one item when time expires.

A solution is to collect data allowing prediction of the number of test-takers likely to have incomplete tests and the number of items they will have answered when time expires. It is then possible to determine the effect of incomplete tests on the accuracy of the individual test and on the overall psychometric characteristics of the test. Using this information, the testing program can specify a minimum number or percent of items required to generate a score.

Establishing Comparability

Within a conventional paper-and-pencil testing program, many mechanisms are employed t insure the comparability of different editions of a test. These mechanisms begin with the assembly of test editions to the same test plan, including statistical properties (based on pretest information) as well as nonstatistical properties of items. These mechanisms end with the process of test score equating that takes place after a test edition has been administered to a large group of test-takers. This equating process compensates statistically for any lack of comparability among test editions.

If a testing program is going to exist with both conventional and adaptive tests or it is making a transition to an adaptive test, but wishes to maintain its score scale, issues surrounding comparability become somewhat different. Studies must be conducted to ensure that the scores obtained from the adaptive test are comparable to those from the conventional test. There are at least two potential sources of non-comparability: the mode of the tests (conventional or adaptive) and the medium (paper or computer). Both can be assessed simultaneously, but if the results indicate the scores are not

comparable, follow-up studies may be required to determine the source of the non-comparability.

The research on comparability has been mixed. Early studies (for example, Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein, 1991) were not promising, but more recent ones (Schaeffer et al., 1993; O'Neill, Folk and Li, 1993) have supported assertions of comparability. The reasons for the inconsistencies are not clear, but could include the quality of the interface between the test-taker and the computer, increased sophistication of computer processors, and higher quality displays.

Developing CAT Pools

There is some guidance on this topic from the theoretical literature on adaptive testing, but almost none of it is helpful. In theory, the most desirable pool of items for the purpose of adaptive testing should have as many items as possible (for security concerns) at all possible levels of test-taker proficiency (for measurement concerns) that match all aspects of the test blueprint (for comparability concerns). While this is a worthy goal to strive for, it is unlikely to be met in practice.

In actuality, we continue to write items in the same fashion that we have developed for paper-and-pencil testing. If the adaptive test is not linked to any existing linear test, there is usually some effort to write items that have a wider range of difficulty while still matching the test blueprint. But the number of items written, pretested, and accepted into a final adaptive test item pool will always be limited by economics. If the adaptive test is linked to existing paper-and-pencil testing, a typical approach to building an item pool is to compose the item pool from items originally pretested for linear tests and also possibly some intact previously

administered linear tests. This insures that there are items in the pool that match the test blueprint and cover an appropriate range of difficulty. But again, the number of items is restricted by current and future pretesting efforts as well as the number of previous forms.

The desire to have as many items as possible is related to the security concerns described in the previous section on controlling item exposure. Adaptive tests are typically administered on a much more frequent basis than conventional tests. A large number of questions are included in the pool from which tests are developed and since different tests are assembled for each test-taker, few, if any, test-takers receive the same test. However, the same item may be administered to many test-takers. Simulations can provide guidance about the extent to which a pool optimizes item exposure, ensuring all items are administered at some base rate while also ensuring that they are not over-exposed.

It may be that a single large pool of many thousands of items gives rise to administrative difficulties in pool management. Certainly such large pools impose certain (expensive) requirements on the hardware and software at testing sites. And certainly a single security break could compromise all available items. For these reasons, it may be useful to subdivide the single large pool into subsets that are more manageable. Again, simulation studies can guide practitioners about the extent to which subpools optimize item exposure and can also lead to the complex and sophisticated rotation of subpools throughout all testing centers as well as within a single testing center. To further enhance security, it is also possible to think of reconstituting subpools frequently so that no single smaller collection of

items from which adaptive tests will be administered will remain intact for a long period of time.

If decisions have been made on the desirable number and size of effective adaptive test item (sub)pools, test specialists are faced with an unfamiliar task -- the efficient development of subsequent (sub)pools of items. This is a different challenge than the development of parallel editions of the same test for linear testing programs, although there are certain similarities. Once initial efforts have produced subsets of items that support adaptive testing, these can serve as blueprints for the construction of subsequent subsets, much as test specialists now produce parallel linear tests. Moreover, depending upon the adaptive testing design used, it may be possible that deficiencies in individual subsets of items can be overcome, at least somewhat, by changing the importance (or weights) attached to some test requirements. Thus the process of building subsequent adaptive testing subpools from a constantly changing very large pool of items may be somewhat easier than the current process of building succeeding parallel editions of the same linear test.

Maintaining CAT Pools

The large item pools supporting operational adaptive testing are unlikely to remain static. Rather they are likely to be in a constant state of flux: newly pretested items may be added to the pools at regular intervals and items may be selected to be removed from pools and disclosed to the public for practice purposes or perhaps to satisfy the requirements of test disclosure legislation. Items with high usage rates may be retired from pools either permanently or temporarily to enhance security. If items are temporarily retired, they may be reintroduced into the pools at some later

date. These types of activities raise all of the obvious requirements of the need for sophisticated database management systems and inventory control systems to manage the day-to-day state of the item pools.

In addition, when adaptive tests are used operationally, procedures will be needed to monitor the functioning of the pool and the items within it over time. For example, it is likely that test questions will need to be removed from the pool on some empirically derived basis. The most obvious choice for identifying items to be replaced is frequency of administration. However, depending on how item exposure is controlled, the most frequently used items may be the "best" in the pool for the particular population of test-takers. If the best items are removed, testing programs may be under pressure to produce high quality items at a faster rate than they have in the past or risk the gradual deterioration of the overall quality of the pool or the need to lengthen the adaptive tests built from a pool of lesser quality (or both).

Furthermore, little is known about how well different kinds of test questions function over repeated, frequent administrations. If items of one type are more memorable than items of another, established rules for item generation and pretesting may need to be revised to produce 'memorable' items at a faster rate than was previously the case or to find ways to make items less memorable.

Developing and Prototyping New Item Types

Although the development of new item types is not specifically related to adaptive testing, the introduction of adaptive testing has greatly stimulated interest in new items types because (1) the computer allows administration and scoring of items that simply could not be administered in paper and/or scored efficiently, (2) the adaptive test is shorter and

additional assessment can occur in the same amount of testing time, and (3) measurement of different constructs incorporating concepts developed based on research in cognitive psychology may be possible using the computer.

The introduction of novel item formats into an existing test can be difficult. Among the issues that need to be considered are placement of the new material so that useful pretest data are collected but performance on the operational test is not impaired, the extent to which the format affects the construct being assessed, whether or not test-takers will be motivated to respond to the new items with the same level of effort as on the more familiar items, and whether the performance of the new item formats will change as test-takers become familiar with them.

Complying With Disclosure Requirements

Many testing programs periodically publish copies of recent tests for test-takers and the public. The tests are typically published after they have been administered to tens or hundreds of thousands of individuals within a specified period of time. This is possible because conventional tests are administered to large groups of people on a single day.

In adaptive testing, it could take years for a single item to be administered to a large number of people and much longer for the same test form (consisting of exactly the same set of questions) to be administered to a large group. Thus, disclosing tests based on the number of people who have taken specific items or test forms is not feasible. It is equally infeasible to disclose tests on a time schedule since every item in the pool may be administered within a very short period of time even if the number of people taking the test is relatively small. A disclosure policy that requires disclosure of all items administered within a short period of time would have

the effect of increasing the rate of disclosure many times. In fact, such a policy would likely result in the disclosure of test questions faster than they can be developed and pretested.

Disclosing tests serves important functions in assuring that test-takers have equal access to materials to familiarize themselves with the test and also serves an accountability function for test sponsors. Thus, it is important that mechanisms for disclosure be developed. An acceptable solution is likely to be one that provides (1) periodic disclosure of entire testing pools and (2) opportunity for individuals to review all or some of the items they took within some reasonable period of time following test administration.

Accommodating Low Volume Tests

To date adaptive testing has been implemented in testing programs that could obtain a large amount of data before the initiation of adaptive testing. The appropriate psychometric model underlying adaptive testing using multiple choice items requires data from large numbers of test-takers to estimate the item statistics used to select questions for administration. It is likely that the appropriate psychometric model required for newly developed item types may require even more data to estimate item statistics. However, there are many low volume testing programs that could benefit from the use of computers to deliver their tests and/or could also benefit from the use of more complex types of items than simple multiple-choice. One option currently available to such programs is computer delivery of their existing, conventional tests. However, if the test is a high stakes one, the security risks associated with frequent computer administrations of the tests are high. New measurement models may be needed to provide a means for the development and administration of computer versions of such tests in a secure environment.

Summary

Following decades of research, a powerful new testing methodology, adaptive testing, has become operational. Adaptive testing provides many improvements to the testing process and enables development of new, innovative assessments while improving service to test-takers, making the process more efficient, and providing improved measurement quality. Operational adaptive testing raises numerous practical and policy-related issues not dealt with during the period of theoretical development. This paper has attempted to discuss a number of those issues, suggest possible solutions to some of them and to identify areas in which additional work may be required.

References

Eignor, D. R., Way, W. D., Stocking, M. L., and Steffen, M. (1993). Case studies in computer adaptive test design through simulation (Research Report 93-56). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lunz, M. E., & Bergstrom, B. A. (1991). Comparability of decision for computer adaptive and written examinations. Journal of Allied Health, 20(1), 15-23.

Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administrative conditions. Journal of Educational Measurement, 31(3), 251-263.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computer adaptive tests. Applied Psychological Measurement, 16(1), 33-40.

Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Mulhstein, A. (1991). Comparability of computer and paper-and-pencil scores for two CLEP general examinations (College Board Report 91-5). New York, NY: College Entrance Examination Board.

McBride, J. R. and Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), New Horizons in Testing, (pp 223-236). New York: Academic Press.

O'Neill, K., Folk, V., and Li, M.-Y. (1993). Report on the pretest calibration study for the computer-based academic skills assessments of The Praxis Series: Professional Assessments for Beginning Teachers

(TM). Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Potenza, M. T. (1994). The Exploration of an alternative method for scoring computer adaptive rests. Unpublished doctoral dissertation. Lincoln, NB: University of Nebraska.

Schaeffer, G., Steffen, M., Golub-Smith, M. (1993). Introduction of a computer adaptive GRE general test (Research Report 93-57). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. Applied Psychology: An International Review, 36, 3/4, 263-277.

Stocking, M. L. (1994). An alternative method for scoring adaptive tests (Research Report 94-48). Princeton, NJ: Educational Testing Service.

Stocking, M. L., and Lewis, C. (1995a). A new method of controlling item exposure in computerized adaptive testing. (Research Report 95-xx). Princeton, NJ: Educational Testing Service.

Stocking, M. L., and Lewis, C. (1995b). Controlling item exposure conditional on ability in computerized adaptive testing. (Research Report 95-xx). Princeton, NJ: Educational Testing Service.

Stocking, M. L., and Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17(3), 277-292.

Swanson, L., and Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. Applied Psychological Measurement, 17(2), 151-166.

Sympson J. B., and Hetter, R. D. (1985, October) Controlling item-exposure rates in computerized adaptive testing, as described in Wainer, et al. (1990).

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., and Thissen, D. (1990). Computerized adaptive testing: a primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. Machine-Mediated Learning, 2, 217-282.